基于CNN-BiLSTM模型的在线医疗实体抽取研究*

■ 陈德鑫[1,2] 占袁圆[1,2] 杨兵[1,2] 谢亚霓[3]
[1] 湖北大学教育学院 武汉 430062 [2] 湖北大学智慧学习研究中心 武汉 430062
[3] 荆门市图书馆 荆门 448000

摘要:[目的/意义]在线医疗信息抽取是实现医疗信息检索、医疗信息推荐、个人医疗健康提醒及警示、疾病诊断、公众健康监控、药物不良反应挖掘等服务的基础环节,而医疗实体抽取则是在线医疗信息抽取的首要工作。本文拟解决传统医疗实体抽取严重依赖于人工特征提取且效率低的问题。[方法/过程]以网络文本为研究对象,首先对医疗实体类型和医疗实体抽取的目标进行描述。将在线医疗文本中的医疗实体抽取任务看作序列标注问题来解决,通过对CNN模型和BiLSTM模型基础理论的探讨,构建基于混合深度学习模型CNN-BiLSTM的医疗实体抽取框架。[结果/结论]通过三组对比实验,验证了本文所使用的CNN-BiLSTM模型在医疗实体抽取任务中的有效性。

关键词:深度学习 卷积神经网络 双向长短记忆模型 医疗实体
分类号:G25
DOI:10.13266/j.issn.0252-3116.2019.12.014

随着公众自我健康管理理念及其医疗信息需求的逐渐改变,网络信息资源中蕴含了大量的医疗相关数据,同时为了推进"智慧健康"的发展,近年来在线医疗信息的相关研究也逐渐成为热点。在互联网普及以及移动App广泛应用的今天,有效的在线医疗信息抽取方法对于满足公众用户医疗信息需求、提高医疗信息服务质量至关重要。在线医疗信息抽取的主要研究方法是基于自然语言处理和文本挖掘,从非结构化或半结构化的医疗数据中挖掘出有价值的信息。这种在线医疗文本是专业医疗文本和网络文本的结合,其与临床文本在语言特征上有本质的差异性。在线医疗的文本中,用户/患者描述事件或表达观点时通常使用日常用语,且包括大量情感和观点类的词汇,给相关信息的检索、统计、挖掘等研究造成了一定的障碍。因此,信息抽取在在线医疗信息价值的挖掘中占有重要的地位。1987-1997年,消息理解系列会议(message understanding conference,MUC)确认了信息抽取具体子任务的划分以及不同子任务的评测体系。1999-2008年,自动内容抽取(automatic content extraction,ACE)评测会议也从不同的角度对信息抽取进行了研究,信息抽取在ACE会议的推动下,其研究深度和广度也在不断加强。根据MUC和ACE会议对信息抽取子任务的界定,医疗实体识别是在线医疗信息抽取工作中的关键过程。目前常用的医疗实体抽取技术分为两类:基于医疗词典和规则的方法以及基于机器学习的方法。基于医疗词典和规则的方法不能对复杂的文本进行高效的处理,只在有限的领域取得显著的成效[1];基于机器学习的方法依赖的传统机器学习模型主要为隐马尔科夫、条件随机场及其调整改进模型,此类模型很大程度上依赖于人工提取的特征,且限定于特定的文本资源[2]。鉴于深度学习在语音和图像处理等领域的良好表现,其在自然语言处理领域的理论研究近几年也越来越受到重视。为了更好地利用未标注的在线医疗文本,使用半监督或无监督的方法进行医疗实体抽取,从深度学习的技术层面出发进行中文在线医疗信息抽取研究,其意义不仅在于对已有的医疗信息抽取理论和方法进行拓展,而且具有重要的实践价值。

本文以网络文本中在线医疗实体抽取为研究目标,通过梳理医疗实体抽取的相关研究理论,结合在线医疗信息子语言特点,明确了在在线医疗实体抽

* 本文系湖北省自然科学基金项目"基于深度学习的网络用户心理健康状态研究"(项目编号:2018CFB315)研究成果之一。
作者简介:陈德鑫(ORCID:0000-0003-2627-6585),讲师,博士,E-mail:202chendexin@163.com;占袁圆(ORCID:0000-0001-8981-6564),本科生;杨兵(ORCID:0000-0002-2774-0282),副院长,教授,博士;谢亚霓(ORCID:0000-0001-8537-7402),馆员。
收稿日期:2018-10-09 修回日期:2019-01-07 本文起止页码:105-113 本文责任编辑:易飞

取任务中引入深度学习的意义。文章首先针对医疗实体抽取过程中的相关问题做了简单介绍,然后阐述了混合深度学习模型的构建过程以及网络文本医疗实体的抽取过程,最后以好大夫在线医疗平台中的文本为实验数据,使用 CNN 模型、BiLSTM 模型以及本文所使用的混合深度学习模型 CNN-BiLSTM 进行医疗实体抽取,探讨 CNN-BiLSTM 模型存在的优势以及需要进一步完善的问题,以期为提高在线医疗实体抽取的效率做出努力,为医疗信息服务奠定良好的基础。

1 相关研究

在常用的医疗实体抽取技术中,基于词典和规则的方法在医疗领域应用范围较为广泛[3]。目前国际上已有多个权威机构维护的医疗词典,例如著名的 ICD -10、UMLS 和 SNOMED CTl0。在对电子病历或临床文本进行医疗实体识别时,早期一般都采用基于医疗词典的方式,医疗领域基于医疗词典进行医疗实体抽取的典型工具有 MedLEE、MedKAT 和 cTAKES 三种。虽然医疗词典是信息抽取中一个非常重要的工具,但是仅使用基于医疗词典的方式不能对复杂的文本进行高效的处理,逐渐地基于机器学习的医疗实体抽取方法受到了越来越多的关注。

基于机器学习的方法对医疗实体进行抽取是将医疗实体抽取任务转化成序列标注问题或者分类问题,基于序列标注的方法进行医疗实体抽取能够考虑临近词的标注信息,因此大部分的研究将医疗实体抽取当做序列标注问题来处理。在基于机器学习的方法中,条件随机场模型(conditional random fields, CRF)是目前进行医疗实体抽取较为有效的方法。龙光宇和徐云利用医疗词典得到医学术语的语义信息,然后通过 CRF 结合这些语义信息来识别疾病命名实体[4]。H. Liu 等采用最新的 CRF 思想,制定合适的特征模板以适应病案实体识别的特点[5]。苏娅等针对医疗问答网站中的疾病问答信息,采用 CRF 模型首次进行了在线医疗实体识别工作[6]。但 CRF 模型特征的选择容易影响结果的准确性且训练代价大,涉及到关键技术和核心模型的突破还存在瓶颈。传统的机器学习方法在对越加复杂、多样的在线文本进行处理时,过分依赖于已标注数据(训练数据),当文本的特征增多时模型的开销会快速增长,且对开放领域的信息进行处理时其召回率会明显下降。为了减少对特征工程的需求,在命名实体识别任务中,J. P. C Chiu 和 E. Nichols 提出了使用 BiLSTM-CNN 架构自动检测字和字符级特征[7]。随着深度学习技术研究的不断深入,其为医疗实体抽取提供了新思路。在信息抽取任务中,杨红梅等基于双向长短记忆模型实现了电子病历文本自动命名实体识别[8]。吴嘉伟等通过稀疏自编码网络模型进行特征再表示提高了信息抽取的召回率[9]。K. Xu 等提出了一种基于双向长短记忆模型和条件随机场的 Bi-LSTM-CRF 医学命名实体识别模型,最后实验的测度超过了许多广泛使用的基线方法[10]。X. Dong 等通过训练双向循环神经网络进行中文电子病历命名实体识别,达到自动提取疾病记录和治疗方法等医学知识的目的[11]。具有层次结构的深度学习技术可以自动学习处理文本时所需要的特征,深度学习模型在医疗实体抽取领域的研究与应用还处在起步阶段,大部分的研究工作还在进行当中,拥有巨大的研究价值和研究空间[12]。目前,深度学习在医疗实体抽取任务中主要是以英文语料为对象,而中文和英文的本质差别使得这些模型并不完全适用于中文领域。本文结合中文在线医疗子语言的特点,根据不同深度学习模型的适用范围和优缺点,提出基于 CNN-BiLSTM 模型的在线医疗实体抽取方法,拟解决传统的抽取方法严重依赖于人工特征的提取、时间成本过高和提取特征有限等问题,如表 1 所示:

表 1 信息抽取方法总结

信息抽取方法	典型的模型	特点
基于词典和规则的方法	MedLEE、MedKAT、cTAKES	优点是准确率较高,但是召回率相对较低
基于传统机器学习的方法	条件随机场模型、隐马尔可夫模型、最大熵模型	优点是在命名实体识别的速度及识别的准确率上有大幅提升,但是严重依赖于人工提取特征
深度学习技术	卷积神经网络、循环神经网络、长短时记忆网络	通过深度学习模型的提取,获得的多重水平的提取特征具有可重复性,可以在相似情境任务中使用。类似于无监督数据集的调用,进而可获得更多可用信息

2 在线医疗实体抽取问题描述

2.1 医疗实体类型定义

医疗实体抽取是进行在线医疗信息抽取的首要工作,即从在线医疗文本中检测并识别出具有特定意义的医疗实体。美国国立卫生研究院(National Institutes of Health, NIH)资助的 i2b2(informatics for integrating biology and the bedside)会议把常见的医疗实体分为 3 类:医疗问题(medical problems, P)、治疗(treatments, Tr)和检查(tests, Te)[13]。借鉴 i2b2 对医疗实体的定义,本文将 i2b2 定义的医疗问题分为疾病(diseases)和症状(symptoms)两类;将 i2b2 定义的治疗分为药物(medicines)和治疗措施(treatment measures)两类。在传统的临床医疗文本和病历文本中,患者、医生和医院信息都是清晰明确的,而在在线医疗文本中用户发布的信息中并没有明确标明患者、医护人员以及医院医疗实体,结合在线医疗信息子语言特点,本文在 i2b2 定义的医疗实体的基础上增加医疗参与者(participants, Par)这一医疗实体类型,并将其分为患者(patients)、医务人员(medical personnels)、医院(hospitals)三类。综上所述,本文中的医疗实体具体分为 8 类:疾病(diseases)、症状(symptoms)、药物(medicines)、治疗措施(treatment measures)、检查(tests)、患者(patients)、医务人员(medical personnels)、医院(hospitals),此 8 类医疗实体的具体描述、标识符号以及示例如表 2 所示:

表 2 在线医疗实体类型及其描述

实体类型		实体类型描述	示例	标识符号
医疗问题(medical problems, P)	疾病(diseases)	医护人员给出的疾病名称或诊断是 ICD - 10 中定义的疾病名、综合征、等术语	破伤风、脑萎缩、阑尾炎、多动症、阿尔茨海默病	D
	症状(symptoms)	患者或医护人员观察得到的患者身体或精神上的异常表现	呼吸困难、精神萎靡、尿频、跟腱断裂、发烧	S
治疗(treatments, Tr)	药物(medicines)	用于治疗或者预防医疗问题的、有目的地调节人的生理机能的物质,包括其通用名、商品名和化学学术名	胰岛素、快克、乙酰氨基酚、云南白药、Cefixme Capsules	M
	治疗措施(treatment measures)	与治疗或者预防医疗问题相关的手术项目或者疗法,包括相关的医疗器械和设备	支架植入术、放射治疗、呼吸机、导尿管等	TM
	检查(tests, Te)	为了诊断医疗问题而施加于病人、体液或样本检查措施或设备	血常规、细菌培养、核磁共振、胃镜、血压等	T
医疗参与者(participants, Par)	患者(patients)	是指患有疾病或具有某些症状的人,是出现医疗问题的个人	某患者、我、我老公、User_ID 等	Pa
	医务人员(medical personnels)	是指从事医务工作中的诊断、治疗、护理等人员的总称,是治疗患者医疗问题的群体	张医生、李主任、护士、屠呦呦等	MP
	医院(hospitals)	是指以向患者提供医护服务为主要目的的医疗机构,是患者治疗医疗问题以及医务人员所从属的机构	武汉省妇幼保健院(省妇幼)、呼吸内科、消化内科等	H

2.2 医疗实体抽取目标

本研究的最终目标是能够准确地识别并抽取出在线医疗文本中本文定义的医疗实体类型。为了能够跟国际医疗实体抽取展现方式相同,借鉴 i2b2 对医疗实体抽取输出文档的要求,定义输入文档和输出文档的格式。本研究对在线医疗文档进行医疗实体抽取是以句子为基本单位进行的,输出的医疗实体抽取结果结构为:

c = “concept text” offset || t = “concept type”　　式(1)

其中 c = “concept text”表示在线医疗文档中识别到的概念实体短语;offset 表示识别到的概念实体短语的起始和结束的句数和字数;t = “concept type”表示本文定义的 8 类医疗实体类型之一。

3 混合深度学习模型 CNN-BiLSTM 的构建

3.1 在线医疗信息子语言特点分析

文本结构、使用的语言以及语言特点对信息抽取所使用的技术和模型有很大的影响。临床文本与在线医疗文本中主要语义类别的差异会对医疗实体类型的定义和抽取产生影响。适用于传统临床文本信息抽取的方法和框架并不适用于在线医疗文本信息的抽取。通过对本文获取的好大夫在线平台的 5 000 篇文档进行分析,在线医疗文本在文档结构、句型结构、标点符号使用、标题使用方面较为混乱和随意,结构性和规范性与临床文本相比较较差。表 3 对临床文本和在线医疗文本的子语言特点进行了总结分析,揭示了在线医疗文本在内容、语言风格、句子结构、词类的使用、书写

习惯、语义类别 6 个方面的子语言特点。这些特点对在线医疗信息的抽取带来了挑战:在线医疗文本是事实性和经验性内容的结合,且在在线医疗文本中不仅使用表示身体部位的形容词,还包含表达用户情感和观点的形容词。根据提供的智能医疗服务的不同,所使用的信息也不同,如何将事实性信息和经验性(主观性)信息进行分离是在线医疗文本处理领域的关注点。例如,对在线医疗文本的情感进行分析时,描述事实型的信息则需要被过滤掉。鉴于在线医疗文本子语言特点及其对文本处理工具和技术的影响,本文将以深度学习模型为技术支撑,构建在线医疗实体抽取方法。

表 3　临床文本与在线医疗文本子语言特点总结

类目	临床文本	在线医疗文本
内容	事实性、证据性信息	事实性、经验性、观点性信息
语言风格	医疗领域专业语言	日常用语,包含部分医学术语
句子结构	句型较短 语法结构不规范	句型较长,结构复杂
词类的使用	使用名词频率相对较高 使用动词频率较低 形容词主要为形容身体不同部位词汇 人称代词为他、她、某患者	使用名词频率相对较低 使用动词频率较高 形容词除了形容身体不同部位词汇,还包含情感和观点类词汇 人称代词丰富
书写习惯	缩写 拼写错误几率低	缩写 拼写错误几率高 对医学术语进行解释 习惯引用/转载其他文本
语义类别	疾病、诊断治疗程序、解剖和特定词	医疗参与者、疾病、症状和药物

3.2　典型的深度学习模型

已有研究中用来进行实体抽取的深度学习模型主要有卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)[14]。CNN 由 6 个部分组成,分别是输入层、卷积层、激活函数、池化层、全连接层、预测层。CNN 能够有效处理自然语言处理中的分类任务,一般采用反向传播法则和有监督训练,通过随机初始化权值的初值对模型进行迭代来减少网络误差,最终当权值集合处于稳定状态时进行收敛。RNN 可以看作是权值共享的多层前馈神经网络,其基本思想是神经网络的隐含层有回边(主要利用序列信息),具有以下两个特征:一是 RNN 模型对序列中的每个元素执行相同的操作,输出结果严重依赖于先前的计算;二是 RNN 能够对历史计算信息进行保存(记忆),并利用历史信息对当前的数据进行辅助计算。RNN 是单向向前传播网络,只考虑了序列元素中的历史信息而没有考虑未来信息,但是通常在文本处理领域,历史信息和未来信息同样的重要,双向循环神经网络(bidirectional recurrent neural network, BRNN)的主要思想为在时间步为 t 时,将序列数据分别从前后两个方向输入到模型中,使用两个隐含层保存的两个方面的隐藏信息进行计算,即既考虑历史序列元素也考虑未来序列元素来进行输出计算。双向长短记忆模型(bi-directional long short term memory, BiLSTM)就是把 BRNN 思想应用到长短记忆模型中,既解决了 RNN 导数消亡问题,也考虑了前后文本关键信息。BiLSTM 在不同的自然语言处理任务中的表现都出色,是本文所构造的用于在线医疗实体抽取模型的重要组成部分。

CNN 和 RNN/LSTM 等深度学习模型都能在文本分析任务中取得一定的成功,但对文本进行序列标注处理时各有优缺点。N. T. Vu 等[15]对 CNN 和 RNN 在文本关系分类上的效能进行了比较分析,认为 CNN 在分本分类任务上的效能高于 RNN,并提出使用 CNN 和 RNN 进行互补建模的思想,即使用 RNN 对句子中所有单词的加权组合进行计算完成抽取关系分类需要的特征短语。Y. Wen 等[16]和 H. Adel 等[17]使用 CNN-RNN 组合模型完成了对长句子的分类。W. Yin 等[18]验证说明了在问答系统中 CNN 的表现性能高于 LSTM。Y. N. Dauphin 等[19]将门控机制加入到 CNN 模型中,以期望加入门控的 CNN 能够对长句进行更好的处理。而 K. Arkhipenko 等[20]将 CNN 与 LSTM 对微博情感分析的性能进行了对比研究,结果表明 LSTM 的性能优于 CNN。

3.3　基于深度学习模型的在线医疗实体抽取方法的构建

通过对文献研究结果进行比较,CNN 能够对文本的字词层级的信息进行学习和表示,但仅能考虑有限范围内的前后字词信息;BiLSTM 能够对文本句子层级的信息进行学习和表示,但其对细粒度的字词特征抽取较差,BiLSTM 模型更擅长处理序列标注问题。基于在线医疗信息子语言特点和不同深度学习模型的适用范围,选择适当的深度学习模型完成在线医疗信息抽取的任务。基于此,本文使用混合深度学习模型 CNN-BiLSTM 对在线医疗文本中的医疗实体进行抽取,整体框架见图 1。医疗实体抽取的本质是对每个字词进行类型标注,因此首先底层使用 CNN 模型对在线医疗文本词汇中表示身体部位、药物命名前后缀、疾病命名前后缀的特征进行学习和表示;其次将 CNN 抽取的字词特征作为 BiLSTM 模型的输入,使用 BiLSTM 模型学习每个字词的历史(前置)信息和未来(后置)信息,进而

陈德鑫，占袁圆，杨兵，等. 基于 CNN-BiLSTM 模型的在线医疗实体抽取研究[J]. 图书情报工作,2019,63(12):105－113.

完成对当前词特征的进一步提取。按照数据的处理过程的流向,CNN-BiLSTM 模型架构可以分为以下几个部分:①文本预处理,对文本进行分词转化成词块;②特征选择,选取能够影响医疗实体抽取效果的特征;③特征向量模块,将选取的特征转化成向量作为输入(此处需要使用 CNN 模型);④将特征向量输入到 BiLSTM 层中进行计算;⑤输出每个词的序列标注概率;⑥选取最大概率的标签作为此词的标注;⑦根据标注结果完成医疗实体抽取。

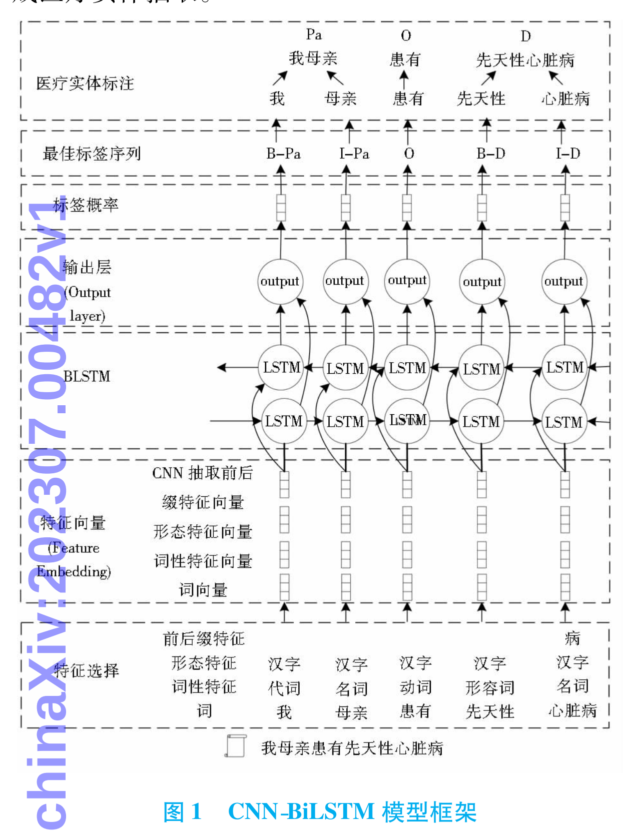


图 1　CNN-BiLSTM 模型框架

4　基于 CNN-BiLSTM 的在线医疗实体抽取过程

4.1　文本预处理

在英文文本序列中可以通过空格对单词进行划分,而中文文本序列是以字为单位,没有明显的分词符,对文本进行处理时需要使用分词工具。中文文本分词的效果能够影响本模型对医疗实体抽取的性能。本文选取的是 Ansj 中文分词工具,此分词工具是在中国科学院 ICTCLAS 中文分词工具的基础上,使用 HMM 算法对其进行了优化和改进,提高了中文分词的准确率。但由于医疗领域词汇的复杂性和特殊性,为了提高分词的准确率,本文将搜狗输入法和百度输入法中的医学词库添加到 Ansj 工具中完成对在线医疗文本的分词。

4.2　医疗实体抽取的特征选择

在特征选取阶段,需要综合一般文本实体抽取时应考虑的特征和在线医疗文本特有的一些特征。本文对在线医疗文本中抽取出的医疗实体特征做了以下 5 种分类:

4.2.1　词特征　词特征是指词语自有的本质特征。

4.2.2　词性特征　中文汉字的词性包括动词、名词、形容词、副词、代词、数词、量词、连词、介词、助词、叹词、拟声词 12 类。在在线医疗文本中患者的经验性描述中经常会出现“觉得”“患有”“吃”“服用”“开了”等动词,在这些动词之后经常会跟着出现疾病名、药物名等名词性医疗实体。词性特征能够为医疗实体的抽取提供线索,本文直接使用采用 Ansj 分词后的词性。

4.2.3　形态特征　形态特征是指当前词语是汉字、数字、英文或其他。英文特征是指当前词中是否包含英文字母,例如检查实体中的“脑 CT”“X 光”等。数字特征是指当前词中是否包含数字,例如高压 120、心跳 90 等词汇。本文直接采用 Ansj 分词后的形态特征。

4.2.4　前后缀特征　词的前后缀特征主要包含本文定义的医疗实体中的后缀词和身体部位指示词两个方面。在英文命名实体识别领域,经常会对词的前后缀特征进行利用,且此方法被证明是有效的。而中文在线医疗文本中的医疗实体也是有一定规律的,例如:疾病实体经常包含病、炎、癌、症、喘等后缀词;症状实体经常包含痛、疼、高、血等后缀词;检查实体经常包含压、查、检、镜等后缀词;药物实体经常包含药、片、素、囊、剂等后缀词;治疗措施实体经常包含疗、术、液、架等后缀词;患者实体在在线医疗文本中通常以某患者或者人称代词和身份性的名词组成;医护人员实体经常包含医生、护士、主任、医师等后缀;医院实体经常以医院为后缀。身体部位指示词是对身体部位进行相关描述的词汇,经常在疾病名称和症状描述中出现。身体部位指示词还包括上、下、左、右等描述方位的词,方位词通常会与身体器官词进行组合。

4.2.5　上下文特征　上下文特征是指句子的语境信息,在 BiLSTM 中既能对历史信息(句子中此词之前的信息)也能对未来信息(句子中此词之后的信息)进行利用,因此本文在模型中并没有把上下文特征放在特征选择和特征向量模块,而是放在 BiLSTM 模型部分进行处理。

4.3　医疗实体抽取的 feature embedding 处理

Feature Embedding 处理模块是对特征选择模块所

选特征进行向量化的过程。

4.3.1　词向量（word embedding）　词向量是把深度学习相关模型引入到自然语言处理领域的一个关键的技术，本文选择使用 Word2vec 进行词向量转换，词向量最终是由一个矩阵表示 $W^{wrd} \in R^{d_w \times |V|}$，其中 V 表示词表的大小，$d^w$ 表示词向量的维度。在此矩阵中，每一列都对应着一个词的向量，即 $W^{wrd} \in R^{d_w}$。使用大规模无监督学习得到的向量作为 $W^{wrd} \in R^{d_w \times |V|}$ 初始值的效果要比使用随机初始化的好，但是国内目前并没有对医疗语料库进行词向量预训练构建的初始值可供选择，因此本文将使用随机初始化的值完成词向量的转化，随机向量维度的初始值为 100 维，向量值从$[-\sqrt{\frac{3}{dim}}, +\sqrt{\frac{3}{dim}}]$范围内随机抽取，其中 dim 是向量的维度[21]。

4.3.2　词性特征向量　中文汉字词性包含 12 类，词性特征则用单维向量表示。例如名词特征向量可以表示为[1,0,0,0,0,0,0,0,0,0,0,0]。

4.3.3　形态特征向量　本文只把形态特征向量分为 4 类：汉字、数字、英文、其他。所以汉字特征向量表示为[1,0,0,0]；数字特征向量表示为[0,1,0,0]；英文特征向量表示为[0,0,1,0]；其他类特征向量表示为[0,0,0,1]。

4.3.4　前后缀特征向量　C. N. D. Santos 和 B. Zadrozny[22]的研究表明 CNN 能够有效地把词汇的前后缀特征抽取出来，并编码成向量形式表示。图 2 为使用 CNN 抽取在线医疗文本词汇前后缀特征的过程图。对于每个词将使用一个卷积层和一个池化层完成特征的抽取。根据 CNN 窗口的大小，每个词会有不同个数的填充向量。此处，CNN 的超参数包括窗口尺寸的大小和输出向量的大小。本文使用随机初始化值，向量值从$[-\sqrt{\frac{3}{dim}}, +\sqrt{\frac{3}{dim}}]$范围内随机抽取，设置维度 dim 为 30。假设输入的样本句子 x 由 N 个词汇组成，$x = \{w_1, w_2, w_3, \cdots, w_N\}$，与每个词汇相对应的 4 种特征分别用 w_i^{kj}（$1 \leq j \leq 4$）表示，其中 kj 表示第 j 特征。$W^{wrd} \in R^{d_w \times |V|}$ 表示词向量矩阵，$W^{kj} \in R^{d^{kj} \times |V_j|}$（$1 \leq j \leq 4$）表示每个特征向量矩阵，其中 V 与 Vj 分别表示词表的大小和各个特征取值区域的长度；d_w 表示词向量维度；d^{wj} 表示每个特征向量维度。最后得到的词向量用 r^w 表示，最后得到的特征向量用 r^{kj} 表示。最终每个词所对应的向量化表示为词向量和特征向量的连接，即：$x_i = [r_i^w, r_i^{k1}, r_i^{k2}, r_i^{k3}, r_i^{k4}]$。

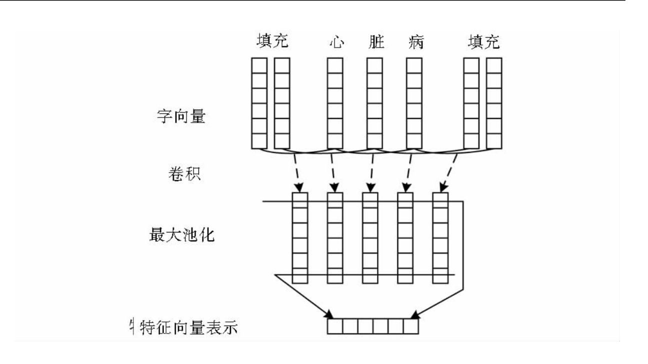


图 2　CNN 抽取词汇前后缀特征过程

4.4　BiLSTM 层

BiLSTM 层能够对句子中的所有文本序列元素信息进行利用，主要为各个词汇之间的相互关系信息，并能将这些信息应用于各个词汇的处理计算中。BiLSTM 的输入即为 feature embedding 的输出。将 feature embedding 层输出的某些位置值随机置成 0，利用 BiLSTM 和前后向计算公式[23]，使用 BiLSTM 从前后两个方向（历史信息和未来信息）输入数据，使得每个词汇的输出都包含了其整个句子中的上下文信息，进而得到 BiLSTM 层第 i 个词对应的输出。

4.5　标签序列输出

本文采用 L. A. Ramshaw 等[24]提出的 BIO 模型对 BiLSTM 的输出序列进行标注，BIO 模型将分类问题转化为序列标记问题。BIO 模型的格式为 B-X，I-X 和 O，其中 B 代表开始（begin）、I 代表中间（internal）、O 代表其他（other）。采用 A. Graves 等[25]提出的模型框架将 BiLSTM 模型中的输出转化成在线医疗实体识别类别标签的评分。

5　基于 CNN-BiLSTM 的在线医疗实体抽取实验

5.1　实验数据集

由于目前国内没有标准的中文在线医疗文本语料库可用于实验研究，因此本文将根据自建的语料库完成对在线医疗实体抽取的任务。以好大夫在线医疗平台为数据源，抽取 5 000 篇在线医疗文档作为本文在线医疗信息抽取的实验数据集。表 4 是对在线医疗实体抽取实验数据的具体描述。将 5 000 篇在线医疗文档按照 3:1 的比例划分为训练集和测试集。训练集数据包含在线医疗文档 3 750 篇，共包含 157 638 句文本。测试集数据包含在线医疗文档 1 250 篇，共包含 43 897 句文本。本文选择深度学习框架 TensorFlow 进行深度学习模型处理，深度学习模型通过 Python 编程

陈德鑫，占袁圆，杨兵，等. 基于 CNN-BiLSTM 模型的在线医疗实体抽取研究[J]. 图书情报工作,2019,63(12):105－113.

语言实现。

表 4　在线医疗信息抽取实验数据描述

数据集	在线医疗文档数量	句子数量
训练集	3 750 篇	157 638 句
测试集	1 250 篇	43 897 句

5.2　在线医疗实体抽取实验结果及分析

实验以抽取好大夫在线医疗文本中的疾病、症状、药物、治疗措施、检查、患者、医务人员、医院 8 类医疗实体为目标。本实验使用随机初始化的值完成词向量的转化，随机向量维度的初始值为 100 维，向量值从 $[-\sqrt{\frac{3}{100}}, +\sqrt{\frac{3}{100}}]$ 范围内随机抽取；CNN 的超参数包括窗口尺寸的大小和输出向量的大小，也使用随机初始化值，向量值从 $[-\sqrt{\frac{3}{30}}, +\sqrt{\frac{3}{30}}]$ 范围内随机抽取。使用 3 750 篇训练数据集完成对深度学习模型参数集的训练，使用训练完成的深度学习模型对 1 250 篇测试数据集进行 3 组医疗实体抽取对比实验：①第一组仅使用 CNN 模型进行在线医疗实体抽取；②第二组仅使用 BiLSTM 模型进行在线医疗实体抽取；③第三组使用本文所选用的 CNN-BiLSTM 模型进行在线医疗实体抽取，将 CNN 抽取的在线医疗文本特有的前后缀特征输入到 BiLSTM 模型完成医疗实体标注。

同样，由于目前国内没有中文在线医疗文本中医疗实体抽取的标准语料库，不能提供可以进行统计的每个文本中正确的医疗实体个数，且由于在本实验中在线医疗文本数量大，对抽取的医疗实体正确总数是通过随机抽样后进行统计计算得到的。本实验从 8 类医疗实体类型中分别随机抽取 50、100、150 个样本点，对每个样本点与原文档进行人工检查对比分析，得出每个医疗实体样本点的平均准确率后再按此准确率计算所抽出的正确的医疗实体的个数。

三组模型进行医疗实体抽取的实验结果如表 5 所示：

表 5　使用 CNN 模型、BiLSTM 模型、CNN-BiLSTM 进行医疗实体抽取的结果

实验模型	医疗实体	文本医疗实体总数	抽取医疗实体总数	抽取正确医疗实体总数	准确率	召回率	F 值
第一组仅使用 CNN 模型进行医疗实体抽取实验	疾病	10 907	9 983	9 487	0.95	0.87	0.93
	症状	29 372	26 724	20 162	0.75	0.69	0.74
	药物	7 086	6 800	5 968	0.88	0.84	0.87
	治疗措施	4 977	3 769	2 968	0.79	0.60	0.74
	检查	6 962	6 632	5 732	0.86	0.82	0.86
	患者	5 962	5 960	5 900	0.99	0.99	0.99
	医护人员	6 984	6 980	6 706	0.96	0.96	0.96
	医院	1 968	1 963	1 902	0.97	0.97	0.97
第二组仅使用 BiLSTM 模型进行医疗实体抽取	疾病	10 907	10 102	9 682	0.96	0.89	0.94
	症状	29 372	27 125	22 098	0.81	0.75	0.80
	药物	7 086	6 918	6 329	0.91	0.89	0.91
	治疗措施	4 977	4 316	3 829	0.89	0.77	0.86
	检查	6 962	6 712	6 112	0.91	0.88	0.90
	患者	5 962	5 962	5 921	0.99	0.99	0.99
	医护人员	6 984	6 981	6 714	0.96	0.96	0.96
	医院	1 968	1 965	1 909	0.97	0.97	0.97
第三组使用 CNN-BiLSTM 模型进行医疗实体抽取	疾病	10 907	10 502	10 313	0.98	0.95	0.97
	症状	29 372	28 125	25 843	0.92	0.88	0.91
	药物	7 086	7 001	6 729	0.96	0.95	0.96
	治疗措施	4 977	4 518	4 196	0.93	0.84	0.91
	检查	6 962	6 854	6 451	0.94	0.93	0.94
	患者	5 962	5 962	5 931	0.99	0.99	0.99
	医护人员	6 984	6 982	6 724	0.96	0.96	0.96
	医院	1 968	1 968	1 912	0.97	0.97	0.97

从表 5 中的实验结果中可以看到第三组基本词向量 + 词性特征 + 形态特征 + CNN 词前后缀特征 + BiLSTM 抽取结果好于第二组仅使用 BiLSTM 模型进行医疗实体抽取的结果；第二组仅使用 BiLSTM 模型进行

医疗实体抽取结果好于第一组仅使用CNN模型进行医疗实体抽取的结果。证明本文所选用的基于CNN-BiLSTM模型能够有效地对医疗实体进行抽取，与其他两组实验相比，CNN-BiLSTM模型能够在疾病、症状、药物、治疗措施、检查5个医疗实体上有良好的表现。尤其在症状和治疗措施两个医疗实体类型上表现更佳，准确率、召回率以及F值都有显著的提高。但在患者、医护人员以及医院三个医疗实体上三组实验抽取效果没有明显区分，三组实验在这三类医疗实体类型上抽取的准确率、召回率和F值都达到了96%以上，甚至患者实体抽取的三组实验的评估指标都为99%，分析其主要原因为在线医疗实体中患者的书写较为规则，即我、我+关系名词、某患者，医护人员、医院两个医疗实体也有相似的特征，因此在此三种医疗实体上的抽取效果较好。

通过实验数据结果发现在识别的8类实体中，除去患者、医护人员、医院三类医疗实体，药物和疾病的识别效果最好，准确率、召回率和F值三个评估指标都远远超过其他类型的医疗实体。其次是检查和治疗措施，抽取效果最差的为症状。主要原因是因为药物名称和疾病名称一般比较固定，并且用户在输入药物和疾病名称信息时格式相对规整。而对症状、治疗措施、检查的识别相对较差，主要原因是用户在对此种类型的医疗实体进行描述时主观随意性太强，如“喉咙有点痒”“医生让我去做检查”等，主观的随意性增加了医疗实体的抽取难度。

6 结语

在Health2.0背景下在线平台中拥有海量的医疗相关数据，虽然存在在线医疗平台中的医疗知识的质量参差不齐等问题。但随着公众自我健康管理意识的改变，越来越多的用户在网络上对医疗内容进行交互，需求增加的同时也产生了大量的有价值的信息。为了更好地向患者、公众以及医护人员提供高质量的医疗服务，高效的在线医疗信息抽取是实现各种医疗信息服务的基础。本文从深度学习的技术层面出发进行中文在线医疗实体抽取研究，旨在提高在线医疗实体抽取的效率，解决传统医疗实体抽取严重依赖人工特征提取的弊端，从而为深度挖掘在线医疗数据价值提供有力的技术支持。本文选用的CNN-BiLSTM混合深度学习模型在在线医疗实体抽取任务中取得了良好的表现，但在构建模型和数据处理的过程中，本文的研究还存在不足之处：①在进行医疗实体抽取时并没有考虑指代消解任务；②本文根据已有的研究对向量的维度进行选择，采用随机的方式对深度学习模型的参数进行处理，但并没有进一步分析不同参数对混合深度学习模型在网络文本医疗实体抽取效果中的影响；③未将在线医疗数据与传统临床数据进行结合使用，提高医疗知识的真实性和准确性。④在本文的概率估计环境中，没有对CTC和CRF等方法进行具体说明和比较。期望在后续的研究中对以上的不足之处进行进一步改进和完善。

参考文献：

[1] 程健一. 中文电子病历实体关系抽取研究[D]. 哈尔滨：哈尔滨工业大学，2016.

[2] WEI Q, CHEN T, XU R, et al. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks[J]. Database: the journal of biological databases & curation, 2016(1):1-8.

[3] 杨锦锋，于秋滨，关毅，等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报，2014，40(8):1537-1562.

[4] 龙光宇，徐云. CRF与词典相结合的疾病命名实体识别[J]. 微型机与应用，2017(21):51-53.

[5] LIU H, QIN X, FU B. The symptoms and pathogenesis entity recognition of TCM medical records based on CRF[C]// Proceedings of the 2017 international workshop on big data and information security. New York: IEEE, 2016:1479-1484.

[6] 苏娅，刘杰，黄亚楼. 在线医疗文本中的实体识别研究[J]. 北京大学学报(自然科学版)，2016，52(1):1-9.

[7] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4(1): 357-370.

[8] 杨红梅，李琳，杨日东，等. 基于双向LSTM神经网络电子病历命名实体的识别模型[J]. 中国组织工程研究，2018(20): 3237-3242.

[9] 吴嘉伟，关毅，吕新波. 基于深度学习的电子病历中实体关系抽取[J]. 智能计算机与应用，2014，4(3):35-38.

[10] XU K, ZHOU Z, HAO T, et al. A bidirectional LSTM and conditional random fields approach to medical named entity recognition[C]// Proceedings of the international conference on advanced intelligent systems and informatics 2017. Berlin: Springer, 2018: 355-365.

[11] DONG X, CHOWDHURY S, QIAN L, et al. Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records[C]// Proceedings of the international conference on E-Health networking, applications and services. New York: IEEE, 2017:1-4.

[12] MIOTTO R, WANG F, WANG S, et al. Deep learning for healthcare: review, opportunities and challenges[J]. Briefings in bioinformatics, 2017,19(6):1236-1246.

[13] i2b2. 2010i2b2/VA challenge evaluation concept annotation guidelines [EB/OL]. [2018－02－18]. https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf.
[14] KOUTNIK J, GREFF K, GOMEZ F, et al. A clockwork RNN [J]. Computer science, 2014:1863－1871.
[15] VU N T, ADEL H, GUPTA P, et al. Combining recurrent and convolutional neural networks for relation classification[C]// Proceedings of NAACL-HLT. Stroudsburg: ACL, 2016:534－539.
[16] WEN Y, ZHANG W, LUO R, et al. Learning text representation using recurrent convolutional neural network with highway layers [C]// Proceedings of Neu-IR '6 SIGIR workshop on neural information retrieval. New York: ACM, 2016:1－5.
[17] ADEL H, SCHUTZE H. Exploring different dimensions of attention for uncertainty detection[C]// Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2017(1):22－34.
[18] YIN W, SCHUTZE H, XIANG B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs [J]. Transactions of the Association for Computational Linguistics, 2016 (4):259－272.
[19] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[C]// Proceedings of the 34th international conference on machine learning. Sydney: The Journal of Machine Learning Research, 2017:1－9.
[20] ARKHIPENKO K, KOZLOV I, TROFIMOVICH J, et al. Comparison of neural network architectures for sentiment analysis of Russian Tweets[C]// Proceedings of the international conference "Dialogue 2016". Moscow: Computational Linguistics and Intellectual Technologies, 2016:1－10.
[21] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C]//Proceedings of the IEEE international conference on computer vision. New York: IEEE, 2015: 1026－1034.
[22] SANTOS C N D, ZADROZNY B. Learning character-level representations for part-of-speech tagging[C]// Proceedings of international conference on machine learning. Beijing: JMLR, 2014:1818－1826.
[23] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Netw, 2005, 18(5):602－610.
[24] RAMSHAW L A, MARCUS M P. Text Chunking using transformation-based learning [J]. Text speech & language technology, 1995, 11(5):82－94.
[25] GRAVES A, MOHAMED A R, HINTON G. Speech recognition with deep recurrent neural networks[C]// Proceedings of IEEE international conference on acoustics, speech and signal processing. New York: IEEE, 2013:6645－6649.

作者贡献说明：
陈德鑫：论文撰写，研究框架和研究方法的提出；
占袁圆：研究内容的补充和完善；
杨兵：研究方法指导；
谢亚霓：论文订正。

Research on Extraction of Online Medical Entities Based on Mixed Deep Learning Model

Chen Dexin[1,2] Zhan Yuanyuan[1,2] Yang Bing[1,2] Xie Yani[3]
[1] School of Education, Hubei University, Wuhan 430062
[2] Smart Learning Center, Hubei University, Wuhan 430062
[3] Jingmen Library, Jingmen 448000

Abstract: [**Purpose/significance**] Online medical information extraction is the basic link to achieve medical information retrieval, medical information recommendation, personal medical health reminder and warning, disease diagnosis, public health monitoring, drug adverse reaction mining and other services, while medical entity extraction is the primary work of online medical information extraction. This paper aims to solve the problem that traditional medical entity extraction relies heavily on artificial feature extraction and the problem of low efficiency. [**Method/process**] Taking network text as the research object, this paper firstly describes the type of medical entity and the target of extraction of medical entity. Online entity extraction task in medical text was considered a sequence labeling problem to solve, the paper discussed the basic theories of BiLSTM model and the CNN model, and built a model based on hybrid deep learning CNN - BiLSTM medical entity extraction framework. [**Result/conclusion**] At last, the effectiveness of the CNN-BiLSTM model in the medical entity extraction task was verified through three comparison experiments.

Keywords: deep learning convolutional neural network bi-directional long short term memory networks medical entities